A CLASSIFICATION OF SCRIPT IDENTIFICATION SYSTEMS

Rumaan Bashir, Kaiser J. Giri, Javaid Iqbal Bhat
Department of Computer Sciences
Islamic University of Science and Technology, Awantipora, J&K
rumaan.bashir@islamicuniversity.edu.in, kaiser.giri@islamicuniversity.edu.in, javaidonnet@gmail.com

Abstract
Vision is an important property of any entity which allows it to perceive the world around it. Computer Vision imitates the human vision by allowing a machine to read and understand images. The images provided to the computer could be that of a document and thus here "Document Image Analysis"becomes a vital task. The field of Document Image Analysis analyses a document image with a primary focus on identifying the chief contents of the document in order to arrive at results which can beinterpreted as per the needs of the users. Script Identification has now evolved to be a key area in the field of Document Image Analysis necessitated due to the fact that a document predominantly contains text. Languages are written using scripts. A lot of work has been reported over the past few decades in this field. Here, an attempt is made to present a detailed classification of script identification systems.

Keywords
Pattern Recognition, Document Image Analysis, Image Processing, Script identification, Classification.

1. OVERVIEW

Computer Science & consequently Information Technology has inexorably become an ingredient of everything that we encompass in the world. It has touched all the realms of life & provides limitless applications in various fields. The solutions provided thereof have now become indispensable for the human kind.They primarily help to abridge the distances across the globe, enable us to save time and reduce the efforts required for any task.

The advent of the internet during the 1970s, the following decades witnessed the extensive growth of active applications; technical diagnosis, global positioning, automatic surveillance, criminology, remote sensing, intelligent office automation, medical imaging, networking, image processing, and pattern recognition leading to the maturity of the field. This development could easily be seen in the rising number of software & hardware products available in the global market. One such application is the *Visual Perception or Vision.*

The world surrounding us is possible to be perceived& understood by humans through the remarkable characteristics of this Vision. To duplicate the effect of human vision electronically the field of Computer vision has evolved(Schalko, 1990). Computer vision is a very complex system. In comparison to computers, humans have the inherent capacity to understand whatever they see including characters & symbols, easily. Computers need to be trained in order to see as humans see. This training is a very complex procedure.

Therefore, the task of computer vision has been simplified to two levels: *lower-level image processing&higher-level image understanding.*

Over the past fifty years, the image understanding skill of a computer has undergone rigorous research (Tang et al., 1952; Pal &Chaudhuri, 2004; Ghosh et al., 2010), but still it is way behind that of the humans. Humans can easily comprehend what they see, even with varied forms & different contents. Humans can read & decipher various text types and pictorial representations in documents, even if they are written / drawn untidily, ornamentally, up-side-down, incomplete or stroked. Sometimes, the content may be distorted in shape & form, but can be still recognized & understood to humans. As far as computers are concerned, this type of ability has not been completely achieved as yet.

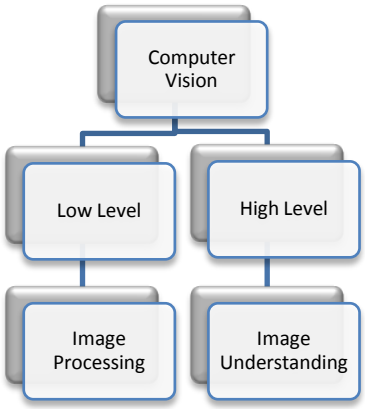


Figure 1: Classification of Computer Vision.

An important area in computer vision is the automatic management of documents. Document is a basic item of manipulation which is used in all fields of life. Documents function asstoring, writing, reading, detailing and processing units/items. Documents have beenoriginally processed by manual procedures which were complicated &time consuming setup. Due to the increase in the use of computers, documents were produced with the help of electronic devices. Following which, there was a shift towards the use of automated & electronic procedures for manipulation of thesedocuments.However, the amount of hardcopy documents was growing colossallyby which the researchers were challenged to develop techniques in order to electronically process the documents and their respective contents in an automated manner. Electronic document management systems have become extremely popular over the past few decades (Pal &Chaudhuri, 2004). These systems are highly applicable in day-to-day life. The automated processing of documents whether machine-generated or man-made emerged as the conception of *Document Image Analysis.*

Keeping in view the requirements of different stakeholders, *Document Image Analysis* is the technique which analyzes a document image for the purpose of extracting itsconstituent parts and contents. The past few decades have witnessed a great deal of investigation, research& development in the much potential

domain of document image analysis which involves acquisition & processing of the documents which are either made manually or made with the help of machines (Bashir &Quadri, 2014, 2015). The processing of these documents primarily focuses on their classification so as to bring efficiency in their storage and use. Document image analysis includes two main areas of *document text processing* and *document graphics processing.*

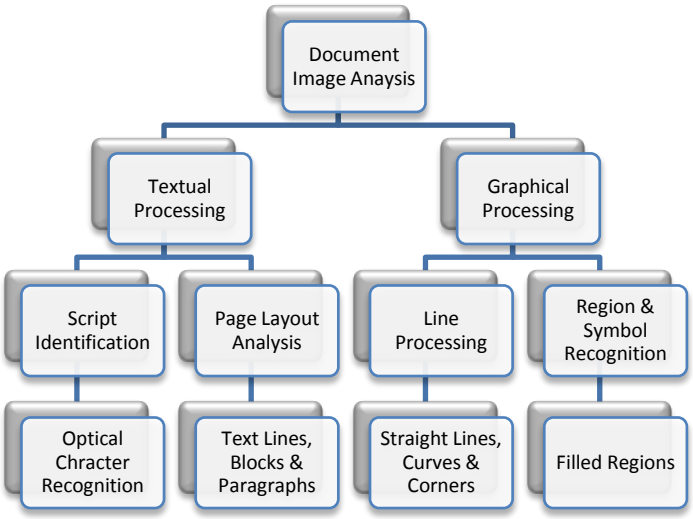


Figure 2: A hierarchy of Document Image Analysis.

Software such as word processors, drawing systems, graphical systems, financial packages computer aided design systems, mark-up languages and the like are extensively used in the production, storage and retrieval of documents. In electronic form, documents can be easily edited, copied and distributed across worldwide networks. In this manner, the documents can be preserved for a long duration of time with minimum space requirement. It can also help to improve the visual quality of ancient valuable writings & legacy works by removing noise which would have been a result of long decay. On the contrary, the hardcopy documents are difficult to preserve and maintain. Storage space requirements for such hardcopy documents are enormous.

As the documents majorly contain text in written form, it calls for analysis and processing of this text. The text processing in document images involves various activities including, the highly important and primary, *identification of the script(s)* in which the documentswereprinted or written.

2. SCRIPT IDENTIFICATION

Document image analysis is recognized for the study of documents manipulation which captures the information content of the particulardocument. Document processing includes detection of *document layout*(O'Gorman 1993; Haralick, 1994) in order to resolve reading/writing order and to distinguish diagrams from text, identification of textual regions, and identification/recognition of the text followed by Optical Character Recognition. This is followed by the processing of diagrams & photographs (Lam, 1994). The identification of text in the form of script(s) has been an active research area for several decades now (Fujisawa, 2008).

The text of a document is preciselycalled a*Script*. Documentsare usually written in one script or can be written in more than one scripts. Every script comprises of an assortment of symbols & characters by means of which it is written. Different languages over the world, which we speak, have been &arestill written using different scripts. A *script* is also represented as a set of graphic signs& symbols e.g. phonetic symbols which are used to write a language. Taking the example of the English language, it is normally written using the Roman script. On the other hand, the Urdu language is written by means ofadapted Arabic script (Toshakhani, 2008).Similarly, the Kashmiri language is also written using modified perso-arabic script. Identification means to ascertain the identity of something. Italso means the act of identifying something. Identification is to relate or link something to a class. Thus, Script Identification means the identification of the script or scripts that documents have been written or printed using, in image form. It is afocal area of intensive study &research in the realm of *Pattern Analysis & Recognition* and *Image Processing*. The electronic image of a document wherein the content is text written/printed in one script or more than one script becomes the input to this procedure.The input is a document image, which is an. Script Identification step is performed prior to the recognition of individual characters by the OCR, as a precursor to OCRing. The output is the decision as to which script has been written in the document image. Various script identification systems have been developed over the time viz. offline & online script identification, machine-written & hand-written script identification, local & global systems, etc. Over the past decades, rigorous research has been carried out to solve this issue in larger context of image processing & pattern recognition. Numerous methods, system architectures and procedures have been presented to deal with the said application diversity. To date, various solutions are being proposed which broadly target to improve accuracy &efficiency of the challenging issues being encountered at various levels. Inautomatic recognition and processing systems, significant progress has been made practically. In the speciality of document analysis, the issue of determining the script(s) of the text in a document image has many other essential applications including indexing,sorting, classification&cataloguing of huge number of images.

India as a diverse country offers many languages & scripts due to the fact that India is a well knownmulti-lingual & multi-script country (Padma &Vijaya, 2009; Padma et al, 2009; Dhandra et al., 2006). In India,

official documents are either written in bilingual scripts or multilingual scripts (Dhandra et al., 2007). The most common script in India is Roman (English) script. Other common script in India is the Devanagari (Hindi) script. The official languages of the specific states are also used to write the documents like Kannada, Tamil, Urdu, etc.

3. CLASSIFICATION

In today's world a different variety of scripts are being used to express the mind in terms of manual writing. There exists an equal essential system which replicates the same concept in the context of computers. It is essential to understand the method & context in which the scripts are represented in this automated system. However, as the multitude of documents grew, whether man-made or machine-made, a need for automatic document processing was also being felt and hence the automatic document management systems. Thus, automatic document processing systems like the document image analysis was formulated, wherein the identification of scripts was a major objective. In literature, there exists a good deal of work which explains how script identification is performed (Ghosh et al., 2010). Many techniques and algorithms have been devised for achieving the best possible method for script identification. The need to summarize the humongous efforts of the work put into script identification in a novel manner inspired us to present the following classification.

Automatic Script Identification can be classified into different types of categories. It depends upon the parameter used for classification. These techniques are based on various parameters which are explained as under:

a. Method of Acquisition
b. Method of Writing
c. Type of Scripts
d. Feature Based
e. Number of Scripts
f. Techniques used for Identification

3.1 METHOD OF ACQUISITION

Data acquisition deals with the mechanisms & tools with which the data is acquired for further processing & recognition. Data is captured by various devices/equipment and is then passed to the computers. The way by which the scripts are written also is an important parameter by which it is determined how script would ultimately be recognized. This applies to the technique of acquiring the document image wherein the text is written in one or more scripts. The document image can be acquired using an online method wherein the text is written and simultaneously its identification is performed. Here, the writing and

identification procedures are executed side by side. In case of the offline method, the document has already been written and the identification is performed after the writing is over. This implies that the writing and identification are performed one after the other and not simultaneously.

- Offline script identification systems: This is a kind of script identification which is applied to a document which has already produced. It is not user the process of production.Offline script identification systems capture the data from paper thorough optical scanners, cameras, and similar equipment. In these methods, the already written text on documents or books are converted into bit patterns with the help of digitizing device like scanners, etc. The identification procedure is then applied on these bit patterns for both machine-written and hand-written text. The bit patterns are normally represented as a matrix of pixels. The size of these matrices depends upon the resolution of the scanning or capturing device. The advantages of this identification method are:
    a. Allows archival data to be processed.
    b. Some application can only work with the help of this method, e.g. postal addresses detection, ticket value detection, cheque sorting, document classification, etc.

    Some disadvantages include:
    a. Costly: it incurs a lot of cost for identification.
    b. Requires more preprocessing: the documents being scanned may lose lot of important information, therefore preprocessing becomes very important step.
    c. This method does not carry the temporal or dynamic information which can otherwise be useful.
    d. This method is not real-time.
- Online script identification systems: This is the kind of script identification where the identificationtakes place when the script is being written. It is a real-time system. Identification procedure is applied in tandem with the writing at the same instance of time. Online script identification systems use digitizers which directly captures drawing as it is done with the pen through identification of strokes, speed, direction, etc. Here, the identification of script is done at the same time, as & when the user enters the text, without waiting. Electromagnetic &electrostatic tablets called the digitizers are usedwhich enable the user to send the position of the tip of the drawing pento the computer at regular intervals. There are pressure-sensitive tablets which perform the same task with a mechanical spacing between the layers of conductive & resistive material. Other technologies include using laser light beams or optical methods. There are various applications of this identification system, e.g. personal devices, mobile devices, PDA's, etc. The online identification systems have many advantages:

a. Real-time system: It captures the dynamic and temporal status of the writing. The information consists of speed of drawing, direction of drawing, number of strokes, etc.
b. Adaptive system: The feedback can be immediately produced and fed back to the recognizer for corrections.
c. Lesser preprocessing: The operations such as smoothing, sharpening, skew correction, slant normalization & feature extraction such as lines, arcs, corners is faster and easier for the pen trajectory than the pixels.
d. Minimal Ambiguity: Optical ambiguity is reduced with the help of pen trajectory.
e. Easier Segmentation: The segmentation becomes easier as the pen positioning and lifting can be clearly identified.
f. It is used to capture hand written text.

Similarly, this recognition system has certain disadvantages:

a. The users require training to use the specialized equipment.
b. It cannot be applied to archival documents.

Figure3: Online script writing using Digital Tablets.

3.2 METHOD OF WRITING

Here, the text written in one or many scripts can be of two types depending upon the fact whether the text is written by a machine or by a human. The difference here lies in the structure of writing as the machine written document are more structured, formatted & contain less noise as compared to the human version. Further, the text is well understood and has relatively less technical or geometric errors i.e. noise. When the input documents to the identification systems, have been created with the help of computerized systems, on good quality paper using modern/latest printing techniques, the identification system yields a good recognizing accuracy. The text in the hand written documents suffers from the issue of variability as it is quite rare that the handwriting of two persons match exactly or with little difference. In addition, the hand-written text normally contains heavy noise due to human errors.

i. Handwritten: If a document contains text script that has been written using hand then handwritten script identification needs to be performed. Here the content of the document is not machine-written. These recognition systems are used to identify such scripts in which text has been written either by *free-hand* or with the help of hand *drawing tools*. Although many techniques have been developed for this purpose but the hand-written script identification is still a challenging task. The complexity of recognition is very high as compared to other recognition applications. This is so because the drawing style, drawing speeds, size of the symbols, aspect ratio, distorted shapes, and physical deformities of human writing cause a lot of variation in the ultimate text.

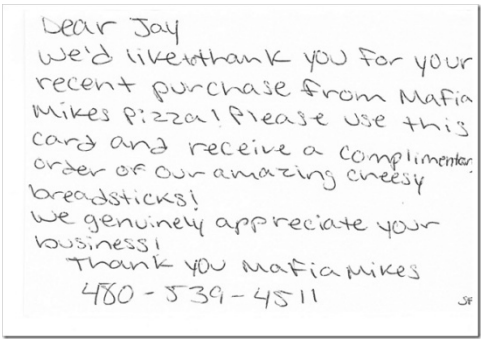


Figure 4: Handwritten Script.

ii. Typeset: If the document contains text script which is printed using aelectronic device e.g. a printer. It also called as machine-written script identification. The input text to such recognition systems may come from books, articles, magazines, journals, and other documents which have been produced by the plotters, printers, etc. The commercially available products for this type of script identification are very much dependent on the source to achieve good recognition rates.

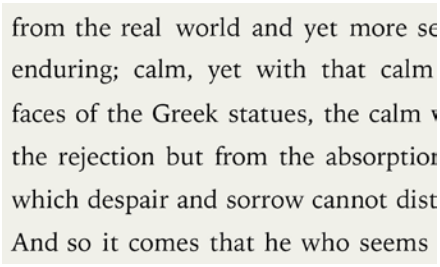


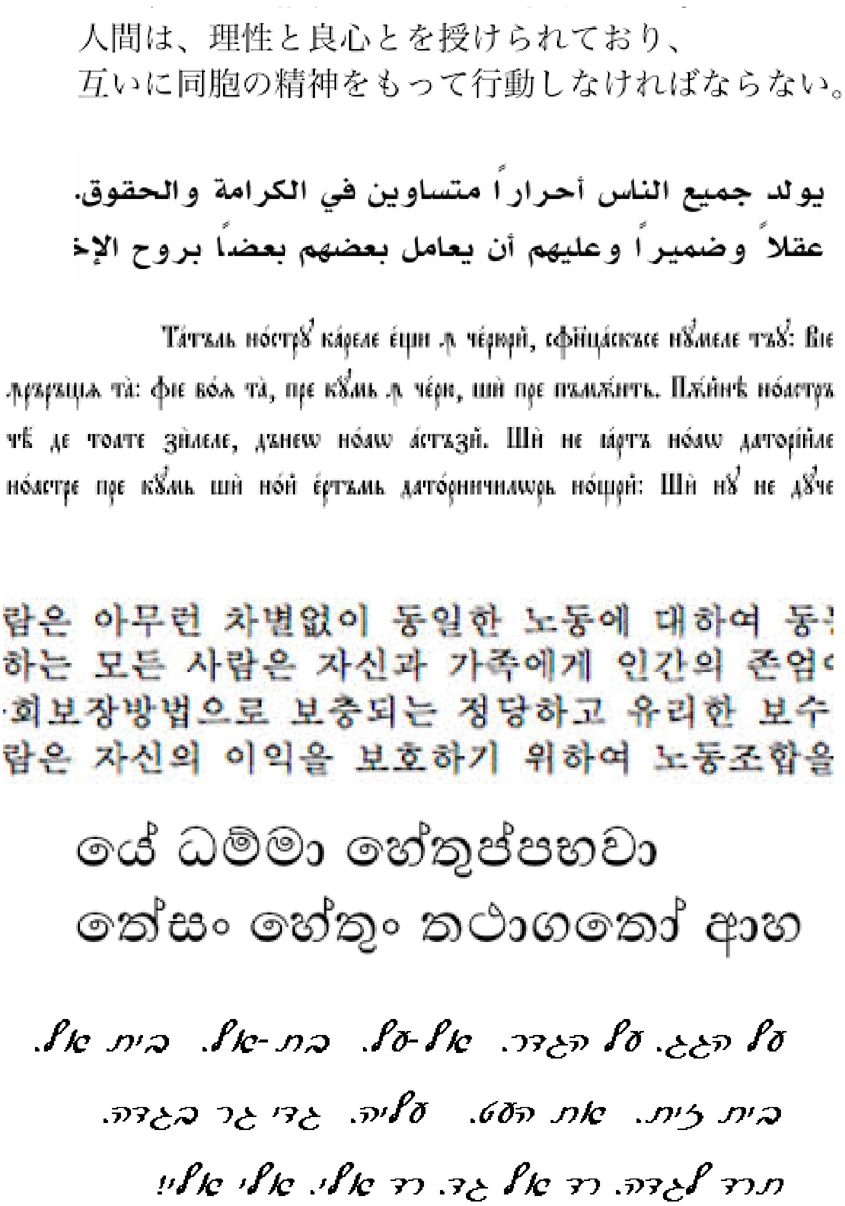
Figure 5: Machine printed Script.

Hand-written script identification systems thus use extensive and complex techniques, pre-processing, segmentation, recognition, etc. & the identification of scripts is comparatively difficult from that of printed/machine-written scripts.

3.3 TYPE OF SCRIPT

The script identification can be applied to the scripts belonging to a particular country. Such scripts are specific to a geographical region are therefore of domestic nature. In some other cases, the script identification is applied to different scripts from different unrelated areas.

i. Foreign/International: Here, the identification of scriptsis performed in a document which contains scripts which do not belong to specific country. A huge set of foreign scripts have been worked out for identification using various procedures including Chinese, Japanese, Russian, Tibetan, Persian, Latin, Sinhala, Han, Korean, Arabic, Persian, Thai, Cyrillic, Sinhala, Hebrew and many more.



ISSN: 2229-6913 (Print), ISSN: 2320-0332 (Online), Web Presence: http://www.ijoes.vidyapublications.com

Figure 6**:** Examples of International Scripts (Sinhala, Hebrew, Arabic, Chinese).

ii. Domestic: Here, the identification of scripts is performed in a document which contains scripts which are used to write languages specific to a country. In India, most of the common regional scripts have been tested for script identification purposes. These include Bangla, Kannada, Gurumukhi, Tamil, Devanagari, Malayalam, Telegu, Marathi, Oriya& Urdu.

बांग्ला	শিবো রক্ষতু গীর্বাণভাষারসাস্বাদতৎপরান্
देवनागरी	शिवो रक्षतु गीर्वाणभाषारसास्वादतत्परान्
गुजराती	શિવો રક્ષતુ ગીર્વાણભાષારસાસ્વાદતત્પરાન્
गुरुमुखी	ਸ਼ਿਵੋ ਰਕ੍ਸ਼ਤੁ ਗੀਰ੍ਵਾਣਭਾਸ਼ਾਰਸਾਸ੍ਵਾਦਤਤ੍ਪਰਾਨ੍
ओडीया	ଶିବୋ ରକ୍ଷତୁ ଗୀର୍ବାଣଭାଷାରସାସ୍ୱାଦତତ୍ପରାନ୍
तमिल	ஷிவோ ரக்ஷது கீர்வாணபாஷாரஸாஸ்வாததத்பராந்
तेलुगु	శివో రక్షతు గీర్వాణభాషారసాస్వాదతత్పరాన్
कन्नड़	ಶಿವೋ ರಕ್ಷತು ಗೀರ್ವಾಣಭಾಷಾರಸಾಸ್ವಾದತತ್ಪರಾನ್
मलयालम	ശിവോ രക്ഷതു ഗീർവാണഭാഷാരസാസ്വാദതത്പരാൻ

Figure 7**:** Examples of Domestic/Indian Scripts.

3.4 FEATURE BASED

This implies the type of feature extracted & employed in order to perform the identification, followed by classification by the script identification algorithm. The algorithm may employ a local feature which may be at the level of a single letter, e.g. the shape of the letter. Contrary to this the feature extracted may be belonging to a whole document called the global feature may used.

i. Local: This sort of script identification is performed for checking the script(s) of a document by only focusing on small area(s) of a document e.g. a character, word, sentence of a paragraph.The features used for identification are restricted to small segments of a document.

ii. Global: This sort of script identification is performed for checking the script(s) of a document by focusing on the whole document in its entirety. The features which are selected for evaluation for performing script identification are spread across the whole document.

3.5 NUMBER OF SCRIPTS

This is a simple concept. Script identification procedure may focus on a document that wherein text is written in only one script, two scripts or many scripts called as unilingual, bilingual and multilingual,

respectively. The complexity of the identification technique also increases subject to the number of scripts in a document. It has three types:

i. Unilingual Identification: Here, the document for identification contains text written in one and only one script. This form of identification is known as unilingual script identification.

ii. Bilingual Identification: Here, the single document for identification contains text written in twoscripts (Bashir &Quadri, 2013). This form of identification is known as bilingual script identification.

iii. Multilingual Identification: Here, the document for identification contains text written greater than two scripts (Bashir &Quadri, 2014). This form of identification is called asmultilingual script identification and is the complex of all.

3.6 TECHNIQUES USED FOR IDENTIFICATION

Based on the techniques that we may use for identification, there are two basic types viz. the Spatial Domain and the Frequency Domain. In the spatial domain, the focus of identification lies on the pixels that comprise the document image. The processing for identification takes place on these picture elements. The spatial domain has further two types. The first type callas the structural domain examines the shape of the writing (script) and the second type uses the statistical methods for the document image. In case of the frequency domain, the document image is transformed to a set of frequencies and then these frequencies are manipulated &evaluated for the purpose of script identification.

i. Spatial-Domain Techniques: For identification purposes, the image pixels of the said document image are thebasic units of manipulation which is called as spatial-domain identification. There are twokinds of techniques in spatial-domain:

- Statistical: When the image pixels of the document image are manipulated in such a manner so as to generate statistical/numerical values which are then used for the identification function, it is called as statistical identification.
- Structural: When the structures of the symbols used for writing the content of the document image, given by the pixels, are used so as toidentifythe script, it is called as structural identification.

ii. Frequency Domain Techniques: For identification purposes, the document image is transformed into frequencies and these frequencies are then evaluated &manipulated to carry out identification then it is called frequency-domain technique.

Based on the aforementioned discussion, the figure given below depicts a novel classification of script identification used for document images.

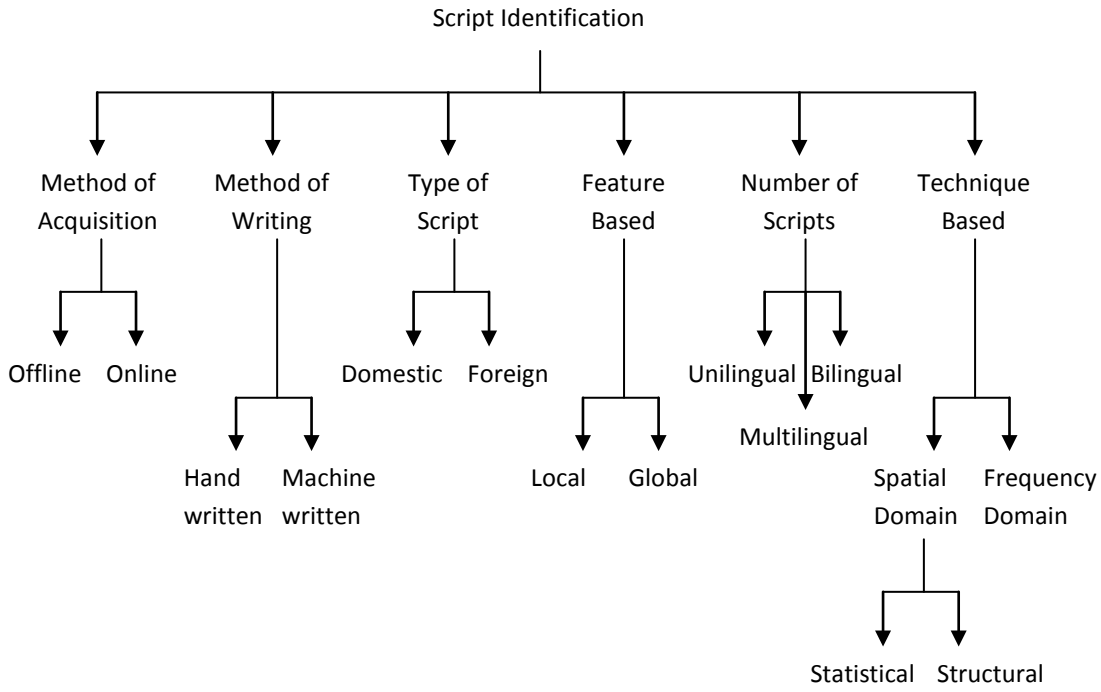


Figure8: Classification Scheme for Script Identification.

4. CONCLUSION

Keeping in view the afore-mentioned, it can be concluded that Script Identification is an area in which a lot of work has been performed and that too using various techniques and methods which could be employed in different scenarios & for different requirements. However, it can also be opined that there could be some scripts used for writing languages which have not been yet been worked out for their automatic identification.

References:

B. V. Dhandra, H. Mallikarjun, Ravindra Hegadi and V. S. Malemath, "Word-wise Script Indentification from Bilingual Document Based on Morphological Reconstruction", IEEE 2006.

B. V. Dhandra, MallikarjunHangarge, RavindraHegadi and V. S. Malemath, "Word-Level Script Identification in Bilingual Documents through Discriminating Features", In Proc. IEEE ICSCN 2007.

Debashis Ghosh, Tulika Dube, & Adamane P. Shivprasad, "Script Recognition – A Review", IEEE, Trans. On PAMI Vol. 32 No. 12 pp 2142-2161 (2010).

Hiromichi Fujisawa, "Forty years of research in character and document recognition- an industrial perspective", Elsevier Pattern Recognition 41 (2008) 2435-2446.

L. O'Gorman and R. Kasturi, "Document Image Analysis", IEEE CS Press, 1995.

L. O'Gorman, "The Document Spectrum for Page Layout Analysis", IEEE Transaction PAMI, Vol. 5 pp 162-173 (1993).

Lam S. , " An adaptive approach to document classification and understanding", In Proc. Intl. Asso. For Pattern Recognition Workshop on Document Analysis Systems, Kaiserlautern, Germany, October 1994 pp. 231-251 (1994).

M. C. Padma and P. A. Vijaya, "Monothetic Separation of Telegu, Hindi and English Text lines from a Multidcript Document", In Proc. IEEE Intl. Conf. Systems, Man and Cybernetics, 2009.

M. C. Padma, P. A. Vijaya, P. Nagabushan, "Language Identification from an Indian Multilingual Document Using profile features", IEEE, Intl. Conf on Computer and Automation Engineering, (2009).

R. Haralick, "Document Image Understanding: geometric and logical layout", In Proc. Intl. Conf. on Computer Vision and Pattern Recognition, 1994, pp. 385-390

R. Schalko, "Digital Image Processing & Computer Vision", McGraw Hill, 1990.

Rumaan Bashir and Quadri, S.M.K., "Entropy based Script Identification of a multilingual Document Image", IEEE Intl. Conf. Computing for Sustainable Global Development (INDIACom), 2014 Page(s): 19 – 23.

Rumaan Bashir and Smk Quadri, "Identifiction of Kashmiri Script in a Bilingual Document Image", In Proc. 2nd IEEE ICIIP, JUIT Shimla, India, (2013).

Rumaan Bashir and S. M. K. Quadri, "Density Based Script Identification of a Multilingual Document Image", I.J. Image, Graphics and Signal Processing, 2015, 1, 1-3 DOI: 10.5815/ijigsp.2014.01.01

S. S. Toshkhani, "Kashmiri Language: Roots, Evolution and Affinity". Kashmiri Overseas Association, Inc. (2008).

U. Pal & B. B. Chaudhuri, "Indian Script Character Recognition: A survey", Elsevier Pattern Recognition 37 (2004) 1887-1899.

Yuan Y. Tang, Seong-Whan Lee and Ching Y.Suen, "Automatic Document Processing: A Survey", Elsevier, Pattern Recognition, Vol. 20, No. 12, pp. 1931-1952.